শূন্য থেকে পাইথন মেশিন লার্নিং

হাতেকলমে সাইকিট-লার্ন

রকিবুল হাসান

হাতেকলমে মেশিন লার্নিং সিরিজ
আইরিস ডেটাসেট প্রজেক্ট

আদর্শ

সূচি

কৃতজ্ঞতা

Data is the new oil. —Clive Humby

ডেটানির্ভর প্রজ্ঞা অর্জন করতে আমাকে সাহায্য করেছে তিনটি অসাধারণ প্রতিষ্ঠান। দিয়েছেন দরকারি প্রশিক্ষন— ডেটাকে বুঝতে তবে ধারণা থেকে নয়; বরং ডেটা থেকে সিদ্ধান্ত নিতে। ১. সিগন্যালস কোর, বাংলাদেশ সেনাবাহিনী, ২. বাংলাদেশ টেলিকমিউনিকেশন রেগুলেটরি কমিশন, ৩. ন্যাশনাল টেলিকমিউনিকেশন মনিটরিং সেন্টার। এরা পুরো ২৭ বছর ধরে আমাকে ডুবিয়ে রেখেছেন টেলিকমিউনিকেশন প্রযুক্তির সঙ্গে। অসাধারণ একটা 'ক্যারিয়ার' তৈরি করে দিয়েছেন আমার। শিখিয়েছেন কীভাবে ডেটাকে যুক্ত করতে হয় সরকারি প্রজ্ঞায়। ধন্যবাদ সরকারি এই তিনটি প্রতিষ্ঠানকে।

টেলিকমিউনিকেশন প্রযুক্তির হৃদয়ের ভেতরটা বুঝতে পারি ট্রেনিং ম্যানুয়াল পড়ে আর 'হাতেকলমে'র অংশগুলো বুঝতে পারি সেনাবাহিনীর সিগন্যালস স্কুলে গিয়ে। সিগন্যালস কোরের নতুন অফিসারদের জন্য এই বাধ্যতামূলক এক বছরের কোর্স আমাকে যতটুকু 'ইনসাইট' দিয়েছে তা ব্যবহার করেই চলছি এখনো। ইলেকট্রনিকস, ডিজিটাল সিগন্যালিং মড্যুলেশন, রেডিও ফ্রিকোয়েন্সি প্রোপাগেশন, অ্যানটেনা ডিজাইন স্কিলসেট শক্ত করে দিয়েছে আমার হবিগুলোকে। একটা সময়ে কমিউনিকেশন ডিভাইস তৈরির জন্য সবাইকে 'এনাবল' করাই ছিল আমার মূলমন্ত্র। ধন্যবাদ সিগন্যালস স্কুল। টেলিকমিউনিকেশন প্রযুক্তির হৃদয়ের অতল বুঝতে। নেশাকে পেশায় পাল্টে দিতে।

ধন্যবাদ বাংলাদেশের টেলিযোগাযোগ সার্ভিস প্রদানকারী কোম্পানিগুলোর অসাধারণ প্রজ্ঞার উদ্যোক্তাদের। যারা আমাকে হাতেকলমে দেখিয়েছেন

ভেতরের ইকোসিস্টেম। এন্ড টু এন্ড। পাশাপাশি ধন্যবাদ সিলিকন ভ্যালির আন্তর্জাতিক ডেটানির্ভর কোম্পানিগুলোতে কর্মরত বন্ধুদের। যারা দেখিয়েছেন কীভাবে কৃত্রিম বুদ্ধিমত্তা পাল্টে দিচ্ছে পৃথিবীর অনেক ধ্যানধারণা। ফর সোশ্যাল গুড। গুগল করে দেখুন, 'মেশিন লার্নিং ফর সোশ্যাল গুড।' যারা প্রতিনিয়ত দেখাচ্ছেন কীভাবে প্রযুক্তি পাল্টে দিচ্ছে ভবিষ্যৎ।

ভালোবাসা স্বাতীকে। আমার স্ত্রী, যিনি আমাকে ছেড়ে দিয়েছেন একটা পুরো ঘর, 'আর এন্ড ডি' ল্যাব হিসেবে। সংসারের সব দায়িত্ব নিজের কাঁধে নিয়ে ছেড়ে দিয়েছেন আমাকে।

শ্রদ্ধা মা-বাবাকে। যারা আমাকে দিয়েছিলেন বইভর্তি বাসা। ছোটবেলায়।

ধন্যবাদ আমার বই, ভিডিও এবং ব্লগ পাঠকদের। যাদের প্রতিটা প্রশ্ন আমাকে শেখাচ্ছে নতুন করে চিন্তা করতে। প্রতিদিন। যারা আমাকে ধারণা দিয়েছেন কেন পুঁথিগত বিদ্যা আর হাতেকলমের মধ্যে এত ফারাক? অসংখ্য ধন্যবাদ আপনাদের। আপনারাই আমাকে উদ্বুদ্ধ করছেন নতুন কিছু শিখতে। প্রতিদিন। টু বিকাম অ্যা রিয়েল প্রবলেম সলভার।

বইটা লেখার জন্য বিশ্বসেরা* ১০টির মতো বইয়ের ধারণা এবং উদাহরণ ব্যবহার করেছি এখানে। পেছনের লিস্টে দিয়েছি সেটা। তবে, আমাকে মুগ্ধ করেছে ডেটাস্কুলের কেভিন মার্কহামের শুরুর দিকের ভিডিও টিউটোরিয়ালগুলো। আমি তার আইরিস ডেটাসেটের ধারণাটা ব্যবহার করেছি এই বইয়ে। ধন্যবাদ কেভিন।

অঙ্ককে পানির মতো লেভেলে আনতে ধন্যবাদ ডাক্সকে। টুয়ার্ডসডেটাসায়েন্স.কম থেকে। একটা ৯৯ পৃষ্ঠার বই এতটা প্রভাব ফেলবে সেটা ধারণায় ছিল না বইটা লেখার সময়। ধন্যবাদ রাউল গ্যারেটাকে, সেই ২০১৩ সালে 'লার্নিং সাইকিট-লার্ন' বইটা লেখার জন্য।

এই বইটা লেখার সময়ে সবচেয়ে বেশি শোনা হয়েছে কেইসি মাসগ্রেভের 'গোল্ডেন আওয়ার' অ্যালবামটা।

---

* অ্যামাজন র‍্যাংকিংয়ে বেস্টসেলার লিস্টে থাকা এই বইগুলো পড়েছেন পুরো পৃথিবীর মানুষ

দ্বিতীয় অধ্যায়

মেশিন লার্নিং কী?

মেশিন লার্নিং জিনিসটা কী?

Being an entrepreneur is like eating glass and staring into the abyss of death. — Elon Musk

আচ্ছা, মেশিন লার্নিং কী?

উত্তর দেওয়ার আগে দুটো জিনিস নিয়ে আলাপ করি।

১. আমাদের ছোটবেলার কথা। ডাক্তারের কাছে গেলেই গুনে গুনে বেশ কিছু জিনিস দেখতেন। শুরুতেই, 'বাবু, জিহ্বা দেখাও।' পাশাপাশি চোখের নিচ অথবা হাত ধরে হার্টবিট দেখতেন প্রায় সব সমস্যার জন্য। এখনকার মতো অত টেস্টিং ফ্যাসিলিটি ছিল না ওই সময়ে। তবুও কিন্তু বলতে পারতেন অনেক কিছু। নির্ভুলভাবে। যেমন—হিমোগ্লোবিন কম না বেশি। কীভাবে?

কীভাবে টেস্ট ছাড়াই প্রায় নির্ভুল 'প্রেডিক্ট' করতে পারতেন ডাক্তার? অভিজ্ঞতা। মানে—মনে রাখা আগের রোগীর ডেটা। যেমন মুখ ফ্যাকাসে থাকলে অথবা চেহারায় ক্লান্তভাব হিমোগ্লোবিনের অভাবের সিম্পটম হিসেবে ধরা হয়। বয়স আর গড়ন অনুযায়ী ওজন কম মনে হলে ডাক্তার ধরতে পারতেন ওই সিম্পটম। সিম্পটমগুলো ছিল ভ্যারিয়েবল। কয়েকটা মিললেই বলতেন হিমোগ্লোবিন কম না বেশি। এই অভিজ্ঞতাটা যখন যন্ত্রকে শেখাব, তখন সেটা হবে মেশিন লার্নিং।

২. ধরুন, আপনি একটি মোবাইল অপারেটর কোম্পানিতে কাজ করছেন। আপনার কাছে গ্রাহকদের বেশ কিছু ডেটা আছে। যেমন ব্যবহারকারী

কোন ধরনের প্যাকেজ ব্যবহার করছেন, দিনে কত কথা বলছেন, গড়ে কত মিনিট করে কথা বলছেন, কখন কখন আপনার সিম খুলে ফেলছেন, ডেটা ব্যবহার কি বাড়াচ্ছেন না কমাচ্ছেন ইত্যাদি বেসিক ডেটা বিশ্লেষণ করতে হবে। এ ছাড়া ওনার মাসিক খরচ, উনি কত দিন ধরে আপনার কোম্পানির পরিষেবা নিচ্ছেন, গত তিন মাস ধরে গড়ে কত মিনিট করে কথা বলছেন, কতটুকু করে ডেটা ব্যবহার হচ্ছে, বাড়তি ডেটা পরের মাসে রোলওভার হচ্ছে কি না, তার হ্যান্ডসেটের মডেল ও দাম, ওই হ্যান্ডসেটটা কত দিন ধরে ব্যবহার করছেন, পোস্টপেইড না প্রিপেইড ইত্যাদি ডেটা আছে আপনার কাছে।

এখন এসব তথ্য থেকে আপনাকে বের করতে হবে—কোন কোন ব্যবহারকারী সামনের মাসে আপনার কোম্পানি ছেড়ে চলে যাবেন?

মেশিন লার্নিং ব্যাপারটাই হচ্ছে যেকোন প্রশ্নের উত্তর ডেটা থেকে দেওয়া যায়। কোন কোন ব্যবহারকারী অন্য মোবাইল অপারেটরে চলে যাবেন, সেটা বের করা যাবে এই আগের ডেটা থেকে। পরের পৃষ্ঠার উদাহরন দেখুন।

ধরা যাক সেলফ ড্রাইভিং কারের কথা। আগে প্রতিটা জিনিস আলাদা করে প্রোগ্রামিং করে দিতে হতো। রাস্তায় মানুষ সামনে পড়লে কী হবে, সামনে ২ ফিট অথবা ১০ ফিট আগে গাড়ি থাকলে কী হবে, রাস্তার পাশ ধরে ডানে যাবে না বামে যাবে, সেটাও প্রোগ্রামিং করে দিতে হতো। ওই প্রোগ্রামিংয়ের বাইরে কোনো ঘটনা ঘটলে, ওই গাড়ি চিৎ পটাং, মানে অ্যাক্সিডেন্ট। কারণ সে ওই না দেখা প্রশ্নের উত্তর জানে না। সে কারণেই মেশিন লার্নিং এল।

অর্থাৎ ‘মেশিন লার্নিং’ হচ্ছে পুরোনো ডেটা থেকে উত্তর পাওয়ার একটা পদ্ধতি। আমরা যেমন শিখি অভিজ্ঞতা থেকে। মেশিনের জন্য অভিজ্ঞতা হচ্ছে পুরোনো ডেটা। ডেটা থেকে জ্ঞান আহরণের এই পদ্ধতিটাই মেশিন লার্নিং।

আবারও বলি, ডেটা থেকে জ্ঞান আহরণের যত পদ্ধতি আমরা ব্যবহার করব, সেটাই মেশিন লার্নিং।

একটু বড় পার্সপেক্টিভ নিয়ে আলাপ করি। আমাদের কাজ হচ্ছে ডেটা নিয়ে। আর সেই ডেটা নিয়ে যত ধরনের কাজ করব সেগুলোকে একসঙ্গে ‘ডেটা সায়েন্স’ বলে। কীভাবে ডেটা থেকে সমস্যার সমাধান করবে, মেশিন লার্নিং হচ্ছে তার একটা অংশ।

'ডেটা সায়েন্স' হচ্ছে শুধু *ডেটা* দিয়ে আপনার কোম্পানির সব সমস্যা মেটানোর একটা বুদ্ধিমান পন্থা। ডেটা দিয়ে আপনার কোম্পানির যত বেশি ইমপ্যাক্ট করা যায়, সেটাই 'ডেটা সায়েন্স'-এর কাজ হবে।

টেক্সটবুকের মেশিন লার্নিং

Say something important rather than say important things. — Daniel Pink.

পৃথিবীর প্রতিটা মেশিন লার্নিং সমস্যাকে আসলে 'তিনটা' কনসেপ্টে ভাগ করা যায়।

ক. শুরুতেই আমাদের শিখতে হবে কীভাবে একটা কাজ বা টাস্ককে (T) সমাধান করতে হবে। ধরা যাক, মোবাইল অপারেটর 'ক'তে কাজ করছেন আপনি। একটা প্রশ্নের উত্তর খুঁজতে দেওয়া হলো আপনাকে, সামনের মাসগুলোতে কোন কোন গ্রাহক ছেড়ে যাবে আপনাদের?

খ. এখন এই কাজটা মেশিনকে শিখিয়ে নিতে অথবা আমাদের নিজেদের করতে কিছু 'অভিজ্ঞতা' বা 'এক্সপেরিয়েন্স' 'E' দরকার। মোবাইল অপারেটরের এই সমস্যার সমাধান আমাদের খুঁজতে হবে কোম্পানির অভিজ্ঞতা থেকে।

এ বিষয়ের কাজ করতে হলে দরকার 'অভিজ্ঞতা', যেটা মোবাইল অপারেটরের কাছে আছে ডেটাসেট হিসেবে। সেই ডেটাসেট থেকে দেখা যাচ্ছে আগে কে কে এই মোবাইল অপারেটর থেকে চলে গেছে। তার চলে যাওয়ার পেছনে কী কী ব্যাপার কাজ করেছে, সেগুলোকে 'ম্যানুয়ালি' বের করার ধারণাগুলো আমাদের জন্য একটা বড় অভিজ্ঞতা। আমাদের মতো মানুষই প্রথমে বের করে দিয়েছে কেন সে চলে গেছে। ফলে ভবিষ্যতে কে কে চলে যাবে তা আগে থেকেই ধারণা করতে পারবে এই মেশিন লার্নিং। বুঝতে পেরেছেন নিশ্চয়ই। আবার বলি, 'অভিজ্ঞতা' হচ্ছে সেই সত্যিকারের পুরোনো ডেটা, যেখানে আমরা জানি কোন কোন কারণে একজন গ্রাহক চলে গেছেন মোবাইল অপারেটর 'ক' ছেড়ে।

গ. এখন 'অভিজ্ঞতা' 'E' থেকে আমাদের 'T' কাজের দক্ষতা কোথায় পৌঁছেছে সেটা জানতে আমাদের দরকার 'পারফরম্যান্স', মানে 'P'। আমরা আসলে কাজটাকে ঠিকমতো করতে পারছি কি না অথবা কতটুকু পারছি, সেটা জানতেই এই দক্ষতার পরিমাপ। আমাদের মেলাতে হবে কতটুকু পারছি ঠিকমতো। সেখানে আমাদের নতুন কিছু যোগ করতে হবে কি না? নতুন কিছু 'মডিফিকেশন' করলে আমাদের উত্তরগুলো ঠিক

আসছে না আরও ভুল হচ্ছে, সেটা বের করার জন্য দরকার আমাদের এই 'P'। 'ক' মোবাইল অপারেটরের কত শতাংশ গ্রাহক আমাদের ছেড়ে চলে যাচ্ছে, সেটার কতটুকু ঠিকমতো ধরতে পারছি, সেটাই আমাদের দক্ষতা। আমাদের প্রশ্নের কত শতাংশ ঠিকমতো কাজ করছে ওই সব গ্রাহককে আগে থেকে 'ক্লাসিফাই' করার ব্যাপারে? 'ক্লাসিফিকেশন' হচ্ছে দুই ভাগে। একজন গ্রাহক, উনি চলে যাবেন, নাকি যাবেন না। আমরা যদি ১০০ জন গ্রাহককে বের করতে পারি যারা চলে যাবেন অপারেটর ছেড়ে, সেখানে বাস্তবে যদি ৯০ জন চলে যান তাহলে আমরা বলতে পারি আমাদের দক্ষতা 'P' হচ্ছে ৯০ শতাংশ।

১. সমস্যা বা প্রশ্নের উত্তরে একটা কাজকে (T) সমাধান করতে হবে
২. কাজ Tকে সমাধান করতে দরকার এক্সপেরিয়েন্স E
৩. কাজটা ঠিকমতো হচ্ছে কি না সেটা দেখতে দরকার পারফরম্যান্স P-এর ব্যবহার। দরকার হলে সেটাকে 'মডিফাই' করে কাজকে আরও ভালোভাবে উতরানোর চেষ্টা করা

অঙ্কে মেশিন লার্নিং ১, প্যাটার্নের ধারনা

The reasonable man adapts himself to the world; the unreasonable one persists in trying to adapt the world to himself. Therefore, all progress depends on the unreasonable man. — George Bernard Shaw

ছবি আর অঙ্কে ডাইভ দেওয়ার আগে ছোটবেলার একটা 'কমন' গল্প বলি। গল্পের উত্তম পুরুষ 'আমি' হয়ত হতে পারেন আপনি নিজেও।

১৯৮০-র দিকের কথা। ছোটবেলা থেকেই আমার একটা অনুসন্ধিৎসু মন ছিল। আশপাশের সবকিছু জানার চেষ্টা ছিল। আশপাশের সবাই ভাবত বড় হয়ে আমি কিছু না কিছু করে ফেলব। ভবিষ্যতে যে 'ডেটা সায়েন্টিস্ট' হিসেবে নতুন ধরনের একটা কাজ আসবে সেটা ওই সময়ে মাথায় ছিল না কারও।

ছোটবেলা থেকেই বিভিন্ন ধরনের প্যার্টান বুঝতে পারতাম আমি। যেমন স্কুলের বন্ধুদের মধ্যে কে কবে ক্লাসে না আসতে পারে, সেটার একটা ধারণা তৈরি করে ফেলেছিলাম। কারণ আমি জানতাম কোন ছাত্র কোন শিক্ষকের ক্লাস ফাঁকি দিতে চায়। আবার শিক্ষকদের মধ্যে কে কে সামনের দিনগুলোতে ক্লাস মিস করতে পারে, সেটারও একটা টেবিল তৈরি করে ফেলেছিলাম পেছনের ক্লাসগুলোর অনুপস্থিতি থেকে। আমার